Earth Science Markup Language: A Solution for Generic Access to Heterogeneous Data Sets

R. Ramachandran*, M. Alshayeb, B. Beaumont, H. Conover, S. Graves, X. Li, S. Movva, A. McDowell and M. Smith

Information Technology and Systems Center
University of Alabama in Huntsville
www.itsc.uah.edu

Abstract **- The Earth Science community is processing and analyzing a large amount and variety of data. These data are generally stored in different data formats, which forces the scientists to spend a significant amount of time in writing specialized, data format specific readers. However, this preprocessing burden can by alleviated by using the Earth Science Markup Language (ESML) to describe the data. ESML is a specialized markup language for Earth Science metadata. Based on the eXtensible Markup Language (XML™), ESML allows data descriptions to be written in a standard fashion, and facilitates the development of data format independent search, visualization, and analysis tools.**

1. INTRODUCTION

The need for structuring documents in a standardized form to facilitate the exchange and manipulation of data dates back to 1960s. The Standard Generalized Markup Language (SGML), a first standardized, structured information technology, emerged as an ISO standard in 1986 [1]. Although extremely powerful, SGML is also too complex for general use. In 1991, the Hyper Text Markup Language (HTML), a subset of SGML, was created by Tim Berners-Lee as a way of marking up technical papers so that they could be easily organized and transferred across different platforms for the scientific community. The idea was to create a set of tags that could be used to mark up a document. The use of these tags would then enable documents to be transferred between computers so that others could render the documents in a usable format. HTML has been the web language of the Internet over the last few years. However, HTML has a significant limitation in its fixed tag set, which prevents the web developers from adding custom tags to HTML so that it can be useful when dealing with data pertaining to a specific application or industry [1]. In 1998, XML™ [2], another subset of SGML, was released. XML has the relative simplicity of HTML and the core benefits of SGML, including extensibility, structure, and validation. XML is a specification for designing markup languages and it defines a standardized text format for representing structured information on the Web. XML defines customized markup languages using Document Type Definitions (DTDs), XML Schema, and other mechanisms. It is expected that XML will play an increasingly important role in the exchange of a wide variety of data on the World-Wide Web. Due to the flexibility of defining customized tags provided by the XML, a number of specialized web-based markup languages have emerged in a last couple of years. These markup languages are defined and used as the standard for information and data exchange on the Web in some specific domains and industries. Examples include Chemical Markup Language (CML)[3], Astronomical Markup Language (AML)[4], and Mathematical Markup Language (MML)[5].

The Earth Science community is processing and gathering huge amounts of data in a variety of different data formats. This large variety of data formats forces scientists to spend a significant amount of time writing specialized data format-specific readers to use the data. Formats for Earth Science data can be as simple as ASCII and Binary, or they can be as complex as Hierarchical Data Format (HDF)[6], and HDF for the Earth Observing System (HDF-EOS)[7]. Fig.1 shows the common situations in dealing with data during typical Earth Science scientific researches. For each new data type, a new program needs to be written to decipher the data for that specific format.

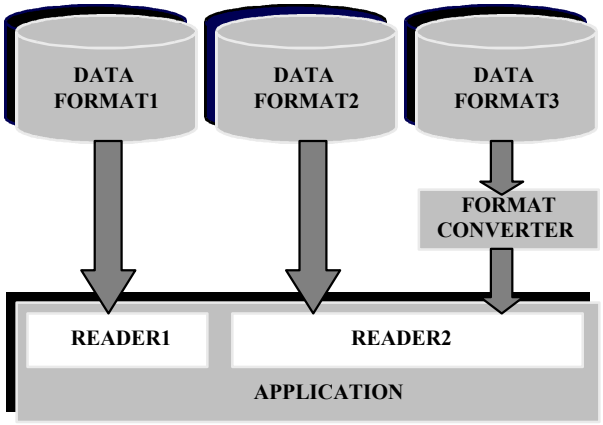

Figure 1. Current state where a new data format requires modifying the application

One solution to the data format problem is to define a standard data format for all earth science datasets. A standardized data format would save scientists a great amount of effort in dealing with the different data formats. HDF-EOS, an extension to HDF, was designed and targeted to be such a standard for the observational data from NASA

satellites. HDF-EOS is a self-describing format based on a logical data model and contains both science data and metadata (a description of the data). The metadata include core metadata that provide information about the content of the dataset, structural metadata that describes the structures in which the data are stored, and archive metadata that facilitates the archiving of data in a database.

Although many software designers and data integrators would prefer that all data be stored in a single format such as HDF-EOS, which would greatly simplify the inter-use, integration, or fusion of disparate data types, the science community has found this solution to be impractical. First, there is a considerable volume of legacy datasets. Converting all of these data to HDF-EOS or another standard format would be exceedingly time-consuming and expensive. Secondly, a particular standard format may not always be the optimal for every data set.

In this paper, the **Earth Science Markup Language (ESML)**, an XML based solution to this data/application interoperability problem, will be described. It is currently being developed with NASA sponsorship at the Information Technology and Systems Center at the University of Alabama in Huntsville.

2. The ESML Concept

ESML allows data descriptions to be written in a separate file in a standard fashion. Each data format can be described in an ESML file, so that an application that can parse an XML™ file will be able to read and interpret the data (Figure 2). The unique feature of ESML is that it not only describes the content and structure of the data, but also attempts to provide semantic information for the end user's application to use. Another advantage is that the effort involved to describe the format of legacy data in ESML is small. Details of ESML schema are described in the next section.

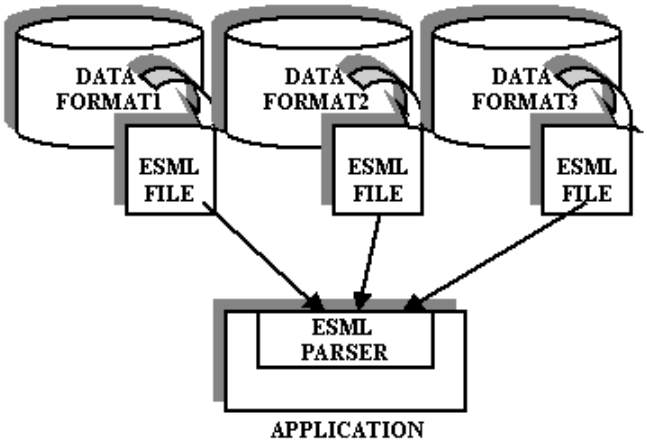

Figure 2. ESML files will make applications data format independent

The ESML files can be published on the World Wide Web (WWW) to facilitate searches via web browsers, illustrated in Figure 3. This will allow any user on the Internet to locate Earth Science data of interest via their web browsers without having to go through a specialized search system.

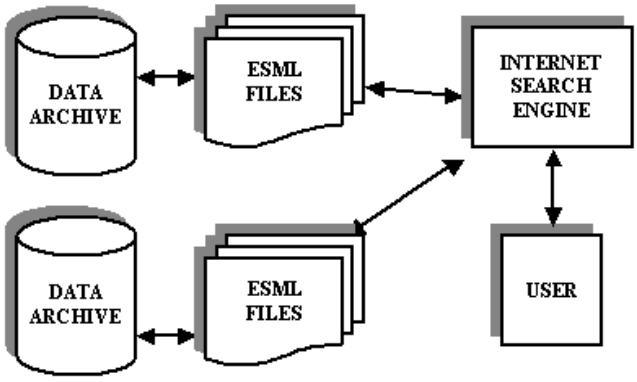

Figure 3. Data search feasible through any Internet search engine

3. ESML schema

ESML provides a means for describing the content, structure and semantic information about different Earth Science data, therefore facilitating development of data format independent access to data products, and applications that do not require data to be in any particular format.

Content metadata describe the contents of a file in human-readable terms. In ESML, content metadata for Earth Science data products are based on the Federal Geographic Data Committee, Global Change Master Directory, the EOSDIS Core System, and other standards. The ESML content metadata contain information such as the producer of the data product and their contact information, spatial and temporal coverage of the data product, data quality, etc., thus enabling search capability. Figure 4 shows part of the ESML content metadata schema.

Syntactic metadata describe the structure of the file in machine-readable terms. HDF-EOS and HDF provide this mechanism for HDF-EOS files. ESML relies on existing standard formats, such as HDF and HDF-EOS, to provide the syntactic metadata when available, but also provides facilities to define these metadata for file formats that are not self-describing, such as ASCII or simple binary files. Figure 5 shows a portion of the ESML Binary schema. The Binary data format can be represented using nested combinations of structures, arrays, If-conditions and ASCII data.

A third level is *Semantic* metadata, which describe the contents of a file in machine-readable terms such that an application can interpret the data in an intelligent manner. For example, this metadata could provide the machine-readable

equivalent of, "This field (described by syntactic metadata) contains latitude values in radians with 0.0 at the equator, increasing northward, scaled by 10000."

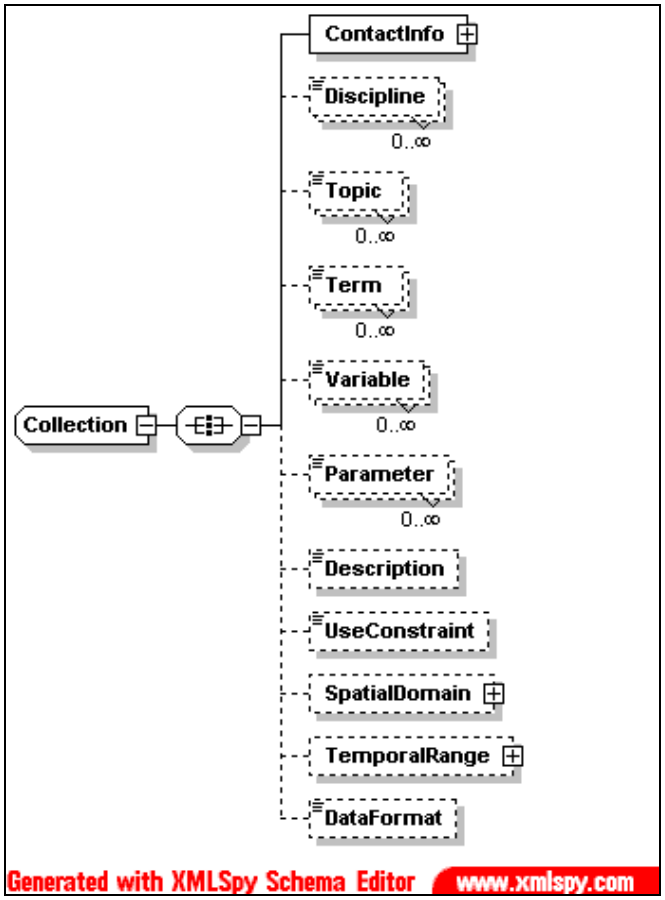

Figure 4: Portion of ESML metadata schema describing metadata that could be used in searches

```
<!-- -->
    <!-- Binary subtype allows any combination of the following types-->
    <xsd:complexType name="binarySubType">
        <xsd:choice maxOccurs="unbounded">
            <xsd:element name="Struct" type="binarraystructType"/>
            <xsd:element name="Array" type="binarraystructType"/>
            <xsd:element name="Datum" type="bindatumType"/>
            <xsd:element name="If" type="binifconditionType"/>
            <xsd:element name="Ascii" type="asciiSubType"/>
        </xsd:choice>
    </xsd:complexType>
<!-- -->
```

Figure 5: Portion of ESML syntactic schema describing possible types of Binary data format structures

At present, HDF-EOS synthesizes many of these metadata by making assumptions about the file and requiring the data producer to use a fixed set of names and formats. This is insufficiently flexible to be of general use. Figure 6 depicts the semantic capabilities available in ESML schema. Each datum type can be labeled as data, attribute (header information), navigation channels (latitude, longitude etc) or time. It also provides a means of specifying equations and mapping.

```
<xsd:complexType name="bindatumType">
      <xsd:sequence>
        <xsd:choice>
            <xsd:element name="Time" type="TimeDef"/>
            <xsd:element name="Latitude" type="LatLonType"/>
            <xsd:element name="Longitude" type="LatLonType"/>
            <xsd:element name="Altitude" type="LatLonType"/>
            <xsd:element name="Data" type="DataDef"/>
            <xsd:element name="Attribute" type="AttributeType"/>
   <!-- embedded *semantic tags* used for labelling information in the files -
->
      </xsd:choice>
      <xsd:element name="Mapping" type="MappingType" minOccurs="0"
maxOccurs="unbounded"/>
      <!-- Mapping information incase the data fields and navigation fields are
not of the same scale -->
```

Figure 6: Portion of ESML semantic schema that allows the user to tag meaning to data fields

An example of ESML for TMI (Tropical Rainfall Measurement Mission - Microwave Imager) Binary data file is provided in Figure 7. The file specifies that there is no explicit navigation information in the file. The data is in a Geographic Projection with the bounds and offsets specified. There are two data fields “SST” and “WIND11” marked by semantic tag <Data/>. These fields are two-dimensional arrays of size 320x1440. The data values in these fields are represented as an Unsigned Integer with bit size of 8.

```
<a:ESML
    <SyntacticMetaData>
        <Binary geoInfo="ByProjection">
            <Projection LowRight_X="360" LowRight_Y="-40"
UpLeft_X="0" UpLeft_Y="40">
                <Geographic latOffset="-0.125" lonOffset="-0.125"/>
            </Projection>
            <BinarySyntactic>
                <Array occurs="320">
                    <Datum name="SST" occurs="1440" type="UInt8"
size="8">
                        <Data/>
                    </Datum>
                </Array>
                <Array occurs="320">
                    <Datum name="WIND11" occurs="1440"
type="UInt8" size="8">
                        <Data/>
                    </Datum>
                </Array>
            </BinarySyntactic>
        </Binary>
    </SyntacticMetaData>
</a:ESML>
```

Figure 7: Example ESML file for TRMM Microwave Imager data in binary format

4. THE ESML LIBRARY

The ESML Library is being designed to provide the end users a simple and easy means to read their data sets and corresponding ESML files. The overall design architecture of the ESML library can be seen in Figure 8. The library provides an easy to use Application Programming Interface. There are three critical components to the library: The XML Parser component reads the XML file and parses it. The Metadata component utilizes the parsed tree to populate its metadata objects.

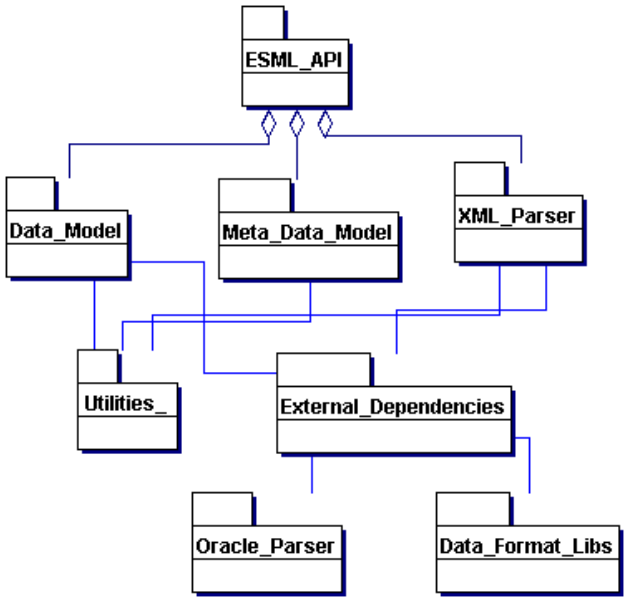

Figure 8. Overall Architecture of ESML library with all its components

Finally, the Data Model component creates data channels as requested by the users to store the data and invokes the read method on the data files. In addition, there are other components that deal with all the utility aspects of the library as well as external dependencies. These external dependencies include shareware XML schema parsers and data format libraries.

5. SUMMARY

The Earth Science Markup Language as presented in this paper provides a means for describing various Earth Science data. ESML is expected to facilitate development of data format independent applications for Earth Science data. The initial version of the schema and more information about this project can be found at: esml.itsc.uah.edu.

ACKNOWLEDGEMENT

The Earth Science Markup Language project has been funded by NASA's Earth Science Technology Office. The ESML development team thanks Karen Moe and Jeanne Behnke at Goddard Space Flight Center, NASA for their guidance of this project.

REFERENCES

[1] Michael Morrison, et al. 2000: XML Unleashed, SAMS Publishing, 960pp.
[2] XML: Extensible Markup Language, http://www.w3.org/XML
[3] CML: Chemical Markup Language, http://www.xml-cml.org
[4] AML: Astronomical Markup Language, http://monet.astro.uiuc.edu/~dguillau/these
[5] MML: Mathematical Markup Language, http://www.w3.org/TR/REC-MathML
[6] HDF-EOS: Hierarchical Data Format for Earth Observing System, http://ivanova.gsfc.nasa.gov/hdfeos
[7] HDF: Hierarchical Data Format, NCSA, http://hdf.ncsa.uiuc.edu